Overview of Open-Source Morphology Development for the Komi-Zyrian Language : Past and Future

Rueter, Jack

The Association for Computational Linguistics
2021

Rueter , J , Partanen , N , Hämäläinen , M & Trosterud , T 2021 , Overview of Open-Source Morphology Development for the Komi-Zyrian Language : Past and Future . in Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages . þý The Association for Computational Linguistics , Stroudsburg , pp. 62 7 Workshop on Computational Linguistics for Uralic Languages 2021 , Syktyvkar , Russian Federation , 23/09/2021 . https://doi.org/10.26615/978-1-954085-82-4_008

http://hdl.handle.net/10138/334834
https://doi.org/10.26615/978-1-954085-82-4_008

unspecified
publishedVersion

Overview of Open-Source Morphology Development for the Komi-Zyrian Language: Past and Future

Jack Rueter
University of Helsinki
jack.rueter@helsinki.fi

Niko Partanen
University of Helsinki
niko.partanen@helsinki.fi

Mika Hämäläinen
University of Helsinki & Rootroo Ltd
mika.hamalainen@helsinki.fi

Trond Trosterud
Norwegian Arctic University
trond.trosterud@uit.no

Abstract

This study describes the on-going development of the finite-state description for an endangered minority language, Komi-Zyrian. This work is located in the context where large written and spoken language corpora are available, which creates a set of unique challenges that have to be, and can be, addressed. We describe how we have designed the transducer so that it can benefit from existing open-source infrastructures and therefore be as reusable as possible.

Дзеныдӧн

Тайӧ гижӧдын сёрни мунӧ канму коми кыв технология йылысь, кӧні сетӧмаӧсь коми морфологиялы помысь-помӧдз автомат. Уджыс сэтшӧм контекстын, кӧні ыджыд гижан да сёрнисикас корпусъяс босьтанног. Та вӧсна чужӧны торйӧн юалӧмъяс, кодлы выль воча кывъяс коланаӧсь. Петкӧдлам, мый эм кыдзи аддзыны колана воча кывъяс. Серпасалам анализатор-автоматлысь сӧвмӧдӧм процесс да вӧзйӧмным ӧтлаӧдны анализаторсӧ паськыдджык восся кодъяса ӧтувтечасӧ-инфраструктураӧ, медым уджыс уналаздоръясын вӧдитчыны.

1 Introduction

This study discusses open-source morphology development, which has greatly benefited from open-source projects most notably achievements attributed to the GiellaLT infrastructure (Moshagen et al., 2014), i.e. Giellatekno & Divvun at the Norwegian Arctic University in Tromsø, Norway. Specifically we discuss the infrastructure for the Komi-Zyrian language. We describe the work done up until now, and delineate some of the tasks we deem necessary in the future. There are features of Komi morphosyntax that need special attention, in regard to both of their linguistic and computational descriptions. This contribution aims to bring that discussion forward, and delineate the current status of the work.

Rueter (2000) describes the initial creation of the transducer, and the work discussed here continues that same undertaking, essentially providing an update of the changes done in the last decade, and a plan for the future. The transducer is available on GitHub for Komi-Zyrian.[1] The nightly builds are available through a Python library called UralicNLP[2] (Hämäläinen, 2019). Easy and efficient access to the traducers and their lexical materials has been the main designing principle, and we consider current approach very successful.

Komi-Zyrian has a growing representation in online corpora. There is a large written corpus that is accessible online[3]; it has been created by FU-Lab in Syktyvkar. The Giellatekno infrastructure provides a Korp implementation (Ahlberg et al., 2013) hosting numerous Uralic Wikipedia corpora, among which Komi can also be found[4]. At the Language Bank of Finland, parallel Bible corpora are available with possibilities for comparing different translations (Rueter and Axelson, 2020). While literary language often reflects astute professional language users, social media provides written language that may be more closely related to the vernacular, this type of Komi is found with minority languages of the adjacent Volga-Kama region[5] and as described in Arkhangelskiy (2019). In a simi-

[1] https://github.com/giellalt/lang-kpv
[2] https://github.com/mikahama/uralicNLP
[3] http://komicorpora.ru
[4] http://gtweb.uit.no/u_korp/#?lang=en
[5] http://komi-zyrian.web-corpora.net/index_en.html

Seventh international workshop on computational linguistics of Uralic languages, pages 62–72
September 23–24, 2021.
https://doi.org/10.26615/978-1-954085-82-4_008

lar vein, a Spoken Komi corpus containing mainly Izhma dialect has been created in a Kone Foundation funded research project (Blokland et al., 2014–2016), and it is also available online for community and research access.[6] Written and spoken language corpora are different in many ways, but together they form a large and representative description of the Komi language. Thereby they both need to be accounted for when the transducer is further developed. Electronic corpora have an important role in the research of Komi in general, and their significance most certainly will only grow when access and practices improve (for discussion about the use of electronic corpora, see Федина, 2019; Чупров, 2018; Блокланд et al., 2014).

There are also numerous dialect materials in Komi, and their progressing digitization gives us access to an increasing number of materials hitherto unavailable in digital format. When this process advances and we inevitably encounter more dialectal texts, we must also consider wider dialectal features of Komi when we develop the transducer.

Additionally, as there are two main Komi varieties with written standards and their dialects, Zyrian and Permyak, we must acknowledge that infrastructures for these languages cannot be developed in isolation, but rather that both language variants must be taken into consideration in different ways (Rueter et al., 2020c). At the same time, the respective written standards have needs for their own tools and resources that are still independent, so the whole question of how to best handle pluricentric language varieties such as Komi still needs additional planning.

The study is structured so that we first describe the work that has been done for modeling the morphosyntax of the Komi-Zyrian language. Then we discuss individual features and their role in the description, and aim to illustrate the types of challenges they present. As we believe that computational modeling of the language is directly connected to the linguistic description itself, we also discuss different phenomena and the ways our description is directly connected to the grammar.

2 Development history of the Komi-Zyrian FST

The FST described here is primarily built by Jack Rueter, beginning with work in the 1990s. The Komi-Zyrian finite-state description began with a trilingual glossary *Ӧшкамӧшка ичӧт кыввор, комиа-англискӧя-финскӧя* (Rueter, 1995), designed for use by Finnish and English speaking students of Komi, without previous knowledge of Russian, to accompany the **коми кыв** '*Komi language*' reader (Цыпанов, 1992), used for instruction in the Universities of Helsinki and Turku. Later, with a scholarship from the Kordelin Foundation, this vocabulary was augmented. First, the extension was intended to complement a second Komi reader by Манова (1994), and then to outline the Komi stem vocabulary of the Komi-Russian dictionary by Лыткин and Тимушев (1961). A large portion of the work done with this dictionary was only possible with the painstaking hours spent by Vera Chernykh. Thus, the approximately 3000-word glossary providing the lexical base for a finite-state description of Komi-Zyrian, presented at **Permistika 6** at the Udmurt State University in Izhevsk, 1996 (published in Rueter, 2000), was extended to over 6000 lexical entries.

In 2004 Trond Trosterud invited Rueter to Tromsø to learn more about the Xerox Finite-state technology (XFST) being implemented at Giellatekno as described in (Trosterud, 2004) and for Komi in (Trosterud, 2004b). Here the Komi transducer and lexicon were to be developed further than before, and to be connected to an infrastructure that was compatible with a larger array of languages.

To summarise some of the new improvements, there were no longer problems with Cyrillic letters requiring representation as conversions from Latin letters. It was now possible to write rules directly addressing elements of the Komi orthography. This direct use of the vernacular in the code may have, in fact, contributed to the belief of the developer that only the normative language needed description. (It was not until many years later that work with other under-resourced languages, such as Mansi (2015–present), Olonets-Karelian (2013–present), Skolt Saami (2015–present) and Võro (2014–present), made it obvious that non-standard words also require description.)

One of the most important items at this point was that the lexicon and morphology were open-source. This meant, in turn, that Komi could be worked on by others and tested in projects. Here, Komi was ideal. The morphology is very concatenative, and the orthography contains only two more letters than the Russian, i.e. problems with some rarer Cyrillic letters could be evaluated and solved.

[6] http://videocorpora.ru

In 2012–2016 Paula Kokkonen worked in conjunction with one of Rueter's projects, where she improved the Finnish translations and inspected the English translations for Komi lexemes. This work significantly increased the coverage of Finnish translations in the multilingual dictionary that was created in this point.

During the period 2012–2021, FU-Lab and Giellatekno collaboration has featured active FST development, including multiple use, and especially improvement in the disambiguation rules and lexical coverage. Morphological analysis is a central component in a modern corpus, and issues such as ambiguity are also always present when FST is used in this context (Оньӧ Лав, 2015, 140). Collaboration may also lead to unforeseeable development. When two infrastructures are aligned, there are often competing priorities. This has also been the case here, i.e. whereas FU-Lab has demonstrated immediate interest in the facilitation of writing, spell checking, dictionaries and corpora for the language community, Giellatekno has pushed for research-related morphological description, analysis and lexica for the research community, but which can, in fact, later be applied to the production of spell checking and other derivative tools. This divergence in priority lead to some duplicate work in morphology.

Helsinki Finite-State Technology (HFST) (Lindén et al., 2013) at Giellatekno with multi-use priorities was pitted against the quick but single-use Hunspell strategies practiced at FU-Lab. Thus, some of the technical complexities on the Giellatekno side had to be simplified so that one set of lexica might be shared. Giellatekno had plenty to gain from the lexical work done at FU-Lab, on the one hand, but it was not able to capitalize on its own sophisticated two-level description as a result of it, on the other. As regards morphophonological descriptions, stem-final variation had to be moved one step away from the initial LEMMA + COLON + STEM + CONTINUATIONLEXICON declaration in the code.

While Jack Rueter has often quickly followed the suggestion of XML maintenance of lexical materials, it has turned out that collaboration pulls away from this write-only-once policy. The more people there are working with one data set, the more documentation required for maintaining mutual working principles. Simple and complex XML systems alike require a working front-end, otherwise, as has been the case here, the workers opt out of the XML database and end up working more on materials that cannot be readily integrated back into the system.

At the moment the XML transformation is not being used in FST development. Instead, other solutions for database implementation are being worked on, see Alnajjar et al. (2020a); Hämäläinen et al. (2021). Only time will reveal which directions of development have contributed the most to the infrastructure.

In 2018-2021, Niko Partanen has been improving the dialectal lexicon coverage of the transducer while conducting his doctoral studies in Komi dialectology. In connection to this work, in 2020-2021, Jack Rueter has improved the coverage of dialectal morphology, specifically taking into account the phenomena found in the Izhma dialect. This work by both of them was done within a Kone Foundation funded research project *Language Documentation Meets Language Technology: The Next Step in the Description of Komi*. The work shows that it is a feasible strategy to improve the analyser so that the work aligns with specific goals and needs of an individual project or dataset. It does create an imbalance in to which degree different dialects are represented, but for a language as large as Komi doing everything at the same time is not possible either.

Mika Hämäläinen's role has been central in building more widely accessible computational infrastructure to access these transducers (Hämäläinen, 2019). In the recent work to create an online editing platform that would allow improved access to the lexical materials, Khalid Alnajjar has been in an irreplaceable position (Alnajjar et al., 2020a). This all shows that managing a transducer for a language like Komi is a multi-partnered operation that calls for wide collaboration between different groups and even infrastructures.

Since 2017, work has been conducted within the Universal Dependencies project to better cover Komi varieties, most recently (Zeman et al., 2021), see also Nivre et al. (2020). There are two Zyrian treebanks (Partanen et al., 2018), and work with Permyak progresses at many levels (Rueter et al., 2020c). Especially in the initial phase of the treebank, building the finite-state descriptions is in a pivotal role, and maintaining interoperability between the FST and treebank development allows very efficient use of both systems. A similar approach has also been systematically used for other languages, such as Karelian (Pirinen, 2019a) and

both Mordvinic languages (Rueter et al., 2020a). Indeed, managing systematic and comparable use of tags and conventions across languages is one of the primary concerns in our work as well, and there have been specific surveys that try to track the progress of different Uralic treebanks (Rueter and Partanen, 2019). We can also mention that the practices described here have also be adopted for the development of Amazon minority language description for Apurinã in Helsinki-Belém (Rueter et al., 2021). In the approach discussed here, this harmonization starts at the transducer level and the documentation therein.

In the context of concrete applications of the Komi FST, we can highlight work by Gerstenberger et al. (2017), where the analyser was integrated into the popular multimedia annotation software ELAN. In addition, the most significant Komi online resource, the National Komi Corpus, contains annotations done with the transducer[7].

Next we describe some of the challenges and important phenomena that have been addressed in various ways when creating the Komi analyser.

3 On describing regular morphology

Komi regular morphology affects word forms in several parts of speech. In addition to verbal conjugation and nominal declension, there is an abundance of regular morpheme-sememe alignment in derivation. Whereas verbal conjugation is, indeed, limited to the indicative (in four synthetic tenses) and imperative moods, the complex noun-phrase head is associated with the categories of number (singular and plural), possessive marking for three persons and two numbers as well as nearly thirty syntactic entity markers or cases. Regular derivation can be observed in aspect, mediopassive and causative marking of verbs, as well as comparative and diminutive marking of nominals. There is a plethora of single-syllable nouns and derivational suffixes, and, at times, the boundary between compounding and derivation becomes obscure.

3.1 Stem variation

The Komi-Zyrian language is known to display a typologically common l-vocalization, which is a process where a lateral approximant is replaced by a labiodental fricative /v/ or labiodental approximant /ʋ/. In the Komi grammaticography this is known as l/v variation. Another comparable stem-alterating phenomena are the paragogic consonants in some word stems. These phenomena can be dealt with in much the same way, as they share a common trigger. Words with l/v or paragogic consonant variation in their stems can be identified on the basis of whether the stem is followed by an vowel-initial suffix, on the one hand, or a consonant-initial suffix (alternatively word boundary), on the other.

In the description of these words it has been suggested that erroneous forms be specifically identified. Special tags indicating the absence of paragogic consonants or substandard realization of the stem-final l/v have been implemented for Komi-Zyrian and reflect parallel tags previously implemented in the FST descriptions of other languages in the GiellaLT infrastructure, Northern and Skolt Saami, Erzya, Moksha, Võro to mention a few.

When we include more dialectal materials in the description, we also have to account for processes where l-vocalization triggers vowel lengthening. There are also secondary types of l-vocalization, influencing stems ending in the sequence /*-el*/, and triggering change /*-ej*/. Currently this is treated at the lemma level, so that the non-standard forms are connected to the standard lemmas, with an additional tag indicating dialectal form or error. Even the dialectal variants where neither types of the variation are met are exceptions from the point of view of the standard language. We have devised a tagging system for various subtypes, but the exact implementation is still being designed and planned further. We discuss in Section 4.2 related challenges in more detail.

3.2 Case

As mentioned above, there are nearly thirty syntactic entity markers or cases associated with complex noun phrases. The distinction drawn here of cases versus derivations lies in the complexity of the noun phrase, i.e. compatibility with the category of number or presence of modifiers has been underlined as a possible boundary (see Rueter, 2010, 74–75; cf. Ylikoski (2020)). If a denominal adverbial derivation does not take adjectival or determiner modifiers, there is no syntactic need to distinguish it from other opaque adverbials. On the contrary, it may be noted, syntactic elements that can take this kind of modifiers should be classified according to their syntactic merits. (The term CASE should not be reguarded as a title of estate but as a useful indication of syntactic class membership.)

[7] http://komicorpora.ru

Here, we will further note that according to the SIL Glossary of Linguistic Terms[8] case is defined as a grammatical category determined by the syntactic or semantic function of a noun or pronoun. If we apply this to a regular morphological description of the Komi languages, we may choose to distinguish between derivational endings applied to simple NP heads and inflectional endings applied to complex NP heads. By distinguishing these two varieties of inflection, we can arrive at a syntactic criterion for classifying different types of inflection, whereas the complex NP, which also takes marking for number, might be readily integrated into the enumeration of nominal modifiers, i.e. cases.

For nearly one and a half centuries, the 16 and 1 dependent cases as defined by Castrén (1844) have represented the canonical cases addressed in grammars of the Komi-Zyrian language. The seventeenth case, the comitative, is addressed as a postposition, but all examples of it show it as integrated morphology in the noun. Некрасова (2000) ('The Modern Komi Language', ÖKK), published in 2000 broke with this tradition by including a set of compounded cases (seven).

The 26 cases shown by the latest Komi grammar, may be further augmented to 29 by introducing the PROPRIETIVE, ABESSIVE and LOCATIVE cases, in *-а*, *-мӧм* and *-са*, respectively. The TEMPORAL in *-ся* might, as a function, be simply attributed to the already existing COMPARATIVE case. Similar questions of case definition have been treated by one of the authors, Rueter (2010), where he regards syntactic entity complexity as sufficient grounds for casehood, (see also Ylikoski, 2020).

Tauli (1956), it should be noted, provides numerous references to researchers dealing with affixes, inclusive derivation and case, there does not seem to be any standards for distinction between case and derivation. The Komi-Zyrian PROPRIETIVE refered to also as a *nomen possessoris* suffix *а*, which occurs as a "comitative" (Tauli, 1956), provides a challenge for the those wishing to distinguish Kom proprietive *-а*, comitative *-кӧд* and instrumental *-ӧн*.

Not unlike the PROPRIETIVE, the ABESSIVE, LOCATIVE and even the temporal function of the COMPARATIVE case are almost entirely limited in use to the adnominal range. The ABESSIVE has a predicative counterpart in the CARATIVE *-мӧг*, while the LOCATIVE has a predicative counterpart in the INESSIVE *-ын*. Perhaps this range distinction has also played a part so-called case classification. The adnominal TEMPORAL marker, however, seems to have no morphological counterpart for use in the predicative.

3.3 Accusative versus object marking

One of the dilemmas in Komi morphosyntax is where to introduce the object of a sentence. Actual non-ambiguous accusative forms are attested for pronouns and other NP heads, but the accusative is not the only case used for indicating the object, the ZERO marker strategy is also used for this purpose. Hence, one might readily speak of object marking with the nominative.

Canonic practice in the Komi grammaticography has been to include the nominative, ZERO form, as an additional accusative case form. If we introduce ZERO as an accusative case marker as well, we, essentially, be introducing ambiguity on the text on the analysis level.

Komi is known for its use of singular possessive suffixes in the accusative for marking different degrees of identifiability; zero, i.e. nominative marking, is also a possibility. When we also have the full syntactic dependency tree, the ambiguity between nominatives and unmarked accusative is resolved, as the object relation is unambiguously marked and connected to the root verb. The current solution in the morphological modeling has been to resolve all unmarked wordforms as nominatives, and to leave the nominative-accusative distinction into a later step of the analysis. None the less, we recognize this is only one of the various ways this can be analysed, and when the full analysis comes, we essentially have all the information to transform the material to match various existing traditions.

3.4 Nominal morpheme ordering

This section will investigate the ordering of morphological constituents typically associated with nominals and convey meaning associated with the categories of number, possession and case.

In initial collaboration with FU-Lab, a singular set of morpheme ordering was adopted for each individual combination of possessor & case marking. Hence, it was determined that the word form *батьӧйлӧн* « бать-ӧй-лӧн 'father.N-PxSg1-Gen' featuring the *ӧй* marking for the first person singular possessor could be distinguished from the possessive suffix *ым* in *гортӧдзым* « горт-ӧдз-ым 'home.N-Ter-PxSg1' on the basis of complementary distribution, i.e. there was no need to label the possessive suffixes as separate entities.

[8] https://glossary.sil.org/term/case

In later development, however, a different issue was observed in which case and possessive formatives might show varied ordering. Although this phenomenon is not as prevalent as in the Meadow and Eastern Mari language (cf. Luutonen (1997)), it did merit recognition and distinction for the facilitation of further resource.

The distinguishing tags strategy implemented for Meadow Mari and Hill Mari has been adapted for use with Komi-Zyrian with two tags. One tag indicates segment ordering where the possessive marker precedes the case marker (+So/PC), and the other indicates the case marker precedes the possessive marker (+So/CP), e.g. *кӧзяиныслань* ← кӧзяин-ыс-лань 'owner.N-PxSg3-Apr' *кӧзяинланьыс* ← кӧзяин-лань-ыс 'owner.N-Apr-PxSg3'.

In addition to this relatively infrequent type of ordering variation of cases versus possessive suffixes, there also appears to be use of the accusative possessive suffix markers for second *-тӧ* and third *-сӧ* person on noun and adjective phrase heads, where the accusative case would not be syntactically compatible. In fact, these same endings are found in connection with other parts of speech as well. It has been maintained that these morphological constituents convey discourse meaning, but there is still much to investigate and establishing tagging practices for these features will contribute to better research materials in the future.

3.5 Numeral derivations

In Komi, numerals are regularly derived to form subgroups in cardinals ZERO, ordinals *-ӧд*, distributives *-ӧн*, iteratives *-ысь*, ordinal iteratives *-ӧдысь* and distributional iteratives *-ысьӧн* (Rueter et al., 2020c). As such, it is often novel or even confounding that we find the syntactic adverbial role found across languages is attributed to a regularly derived adverb *кыкысь* 'twice' on the Komi side, on the one hand, and a noun phrase *fifty times* 'ветымынысь', on the other.

Like other adnominal modifiers, it should be noted, numerals may also be promoted to NP head position in instances of contextually motivated ellipsis.

4 Development plan

We have recently moved into primarily data-driven development practice for Komi, where new lexicon and morphology is described primarily based on gaps we find through analysed language materials. At the same time we have developed further tests to check the validity of the output, and in the long term these approaches naturally will live on in parallel. Needless to say, using more natural texts has also forced us to take into account more spoken language and dialect phenomena, which moves the work into quite new directions, which we have already discussed partially above.

After reporting our experiments with the written corpus data, we discuss our plan to integrate the dialectal materials and tags better to the currently discussed Komi analyser.

4.1 Developing on the basis of unrecognized words and word forms

From a corpus of 1,415,210 unique word forms (2020-11-11) 520,180 were not recognized by the analyzer. Aside from the Russian words, apparently from quoted text, and words written entirely in upper case, the most frequent words not to be recognized by the FST seem to all involve hyphens. The use of hyphenation is best illiustrated by *Рытыв-Войвыв* the preposed modifier for direction 'north northwest' (1377 times), a drawn out pronunciation *Но-о* 'Well-l' (1177 times), and the orthographic practice of adding *-мӧд* 'another' in *здук-мӧд* 'yet another moment' (942 times).

Since over a third of the unique word forms had gone unrecognized, a strategy was developed for improving the model. This would be carried out for nominals initially and subsequently verbs. As described below, a very large portion of unrecognized forms involved various plurals. How they were dealt with is described below, as it illustrates well the challenges we have encountered and their possible solutions.

In the Komi-Zyrian morphology there are two separate plural markers associated with nominal declension. One is the NP plural marker *яс* and the other is the copula complement plural marker *ӧсь*. 20604 unrecognized word forms ended in *яс*, and in 11441 of these the plural marker was preceded by a Cyrillic hard sign ъ. This number was was further delimited by removing all instances of hyphenation and *v* followed by Cyrillic hard sign and word-final *яс*. Where the hyphen may have meant compound words for simple hyphenation in the text, the removal of *v* meant we could automatically avoid the problem of determining whether the word stem contained the notorious *l/v* variation or not. Our resulting figure was 8766.

After entering 15,101 new stems the number of unrecognized unique word forms dropped to 422,227, which was nearly a nineteen per cent improvement over the previous 520,180. In the future we plan to go further through the frequency list of unknown word forms and improve the analyzer so that individual yet frequent phenomena is adequately described and addressed.

4.2 Treatment of dialectal elements

Currently the FST is designed so that dialectal elements are recognized, but they come with an additional error or dialect tag which prevents them being suggested in tools such as spellcheckers. We have also experimented with approaches where Zyrian, Permyak and Russian analysers are run on top of one another, so that unknown forms may be captured by one of the systems with appropriate language tags returned. Since some Zyrian dialectal phenomena is also present in Permyak standard language, already this solution helps to improve the coverage.

Eventually, however, we consider it important that the analyser could capture nuances of individual dialects. In principle this could be accompanied with dialect specific tags, but this approach is also problematic. Many of the features are not strictly found in singular dialects, but cover larger regions. At the same time the speech of any individual is not necessarily limited to any specific variety. Moreover, we believe that further research in Komi dialect isoglosses may be necessary to exactly point for each feature where they definitely occur. Some rough areal boundaries, however, are well known and clear cut, which would make some areal tags potentially useful.

Features that currently are not included are especially those found from southern and eastern Zyrian dialects, mainly because nobody has attempted to use an FST with those varieties yet. We must also recognize that Permyak and Zyrian dialects overlap in their features in various ways, and especially the creation of infrastructure that handless all Komi varieties and both standards remains a challenge.

5 Future directions and Conclusions

In recent years many neural network based approaches have been becoming popular and also shown good results. In a recent study by Pirinen (2019b) the neural models were better than the traditional rule-based approaches for Finnish. Our team is always following new developments of the field, but we also believe that different approaches can be successfully combined.

We already see studies emerging where a neural network has been used to learn to generate predictions from an FST (Hämäläinen et al., 2021). Their research is also used the Komi-Zyrian FST presented in this paper. The results were promising and we are eager to see how this ideology of using neural networks and rule-based systems side by side rather than as competing systems plays out in the future. For the NLP pipeline of Komi the most important new developments will be connected to improvements in the dependency parsing side of the analysis, ideally in connection to automatic and rule-based methods of disambiguation. Komi Constraint Grammar has currently focused to disambiguation, and the tagging and parsing sections are largely missing. It remains to be seen what kind of an approach will be the most successful here. At the same time Komi Universal Dependencies treebanks have started to be large enough that their further modeling with deep learning starts to be an attractive and possibly fruitful task.

Komi texts are also present in many different orthographies, and taking all of them into account is a large and important task (Rueter and Ponomareva, 2019). Since the corpora of Latin Komi texts are also now available[9], the future for these lines of research is exciting and promising. This also connects to various transcription systems used in linguistic publications and text collections: these materials should be republished in the contemporary orthography in order to make them maximally useful for the language communities themselves.

Yet another future task is to provide access to the multilingual Komi lexicon the FST is based on in a form that is truly accessible and openly available. One solution could be to use online dictionary editing platforms, which are strongly linked to the FST development work, and thereby benefit it directly (Alnajjar et al., 2020b). These lexicons have already been published in Zenodo (Rueter et al., 2020b), and already their earliest version has been published in print (Rueter, 1995). Thereby the work described here in various ways continues an already 25 years old progress at morphological modeling of the Komi language, and explores new ways to connect various threads of existing work to one another, especially in ways that takes into account the tech-

[9] http://latina.komicorpora.ru/

nological and practical changes that these decades have shown. We believe this line of investigation of the Komi language will boldly continue the next 25 years, but also hope the reports of how the work progresses will become even more regularly.

We also foresee that further development of the Komi FST will bring new tools to benefit both the public and research communities. Such might be machine translation, on the one hand (Tiedemann, 2021), and translation studies, on the other (cf. Цыпанов, 2021). This, of course, does not close the circle, but merely the ever continuous spiral of development.

Acknowledgements

As described in the study above, this work on Komi has been funded by Alfred Kordelin Foundation and Kone Foundation.

References

Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, and Jonatan Uppström. 2013. Korp and Karp–a bestiary of language resources: the research infrastructure of Språkbanken. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 429–433.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2020a. On editing dictionaries for Uralic languages in an online environment. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 26–30.

Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020b. Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. *arXiv preprint arXiv:2012.02578*.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140.

Rogier Blokland, Vasily Chuprov, Maria Fedina, Marina Fedina, Dmitry Levchenko, Niko Partanen, and Michael Rießler. 2014–2016. Spoken Komi Corpus. The Language Bank of Finland version.

M.A. Castrén. 1844. *Elementa Grammatices Syrjaenae*. Ex officina typographica heredum Simelii, Helsingforsiae.

Ciprian Gerstenberger, Niko Tapio Partanen, Michael Rießler, and Joshua Wilbur. 2017. Instant annotations: Applying NLP methods to the annotation of spoken language documentation corpora. In *International Workshop for Computational Linguistics of Uralic Languages*, pages 25–36. The Association for Computational Linguistics.

Mika Hämäläinen, Khalid Alnajjar, Jack Rueter, Miika Lehtinen, and Niko Partanen. 2021. An online tool developed for post-editing the new Skolt Sami dictionary. In *Electronic lexicography in the 21st century (eLex 2021). Proceedings of the eLex 2021 conference*, pages 653–664, Czech Republic. Lexical Computing CZ s.r.o.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021. Neural Morphology Dataset and Models for Multiple Languages, from the Large to the Endangered. In *Proceedings of the the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. HFST a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 53–71. Springer.

Jorma Luutonen. 1997. *The Variation of Morpheme Order in Mari Declension*, volume 226 of *Suomalais-Ugrilaisen Seuran Toimituksia*. Suomalais-Ugrilainen Seura, Helsinki.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era".

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Tommi A Pirinen. 2019a. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Tommi A Pirinen. 2019b. Neural and rule-based finnish NLP models—expectations, experiments and experiences. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 104–114.

Jack Rueter. 2000. Hel'sinkisa universitetyn kyv tujalys' Iżkaryn perymsa simpozium vylyn lydd'ömtor. In *Permistika 6 (Proceedings of Permistika 6 conference)*, pages 154–158.

Jack Rueter. 2010. *Adnominal person in the morphological system of Erzya*. Number 261 in Suomalais-ugrilaisen seuran toimituksia. Suomalais-Ugrilainen Seura, Finland.

Jack Rueter and Erik Axelson. 2020. Raamatun jakeita uralilaisille kielille: rinnakkaiskorpus, sekoitettu, korp [tekstikorpus].

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies treebank. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020a. Open-source morphology for endangered Mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. The Association for Computational Linguistics.

Jack Rueter, Paula Kokkonen, and Marina Fedina. 2020b. Komi-zyrian-to-x dictionary work. Zenodo data repository, version 0.5.1.

Jack Rueter and Niko Partanen. 2019. Survey of Uralic Universal Dependencies development. In *Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019) Proceedings*. The Association for Computational Linguistics.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020c. On the questions in developing computational infrastructure for Komi-Permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25.

Jack Rueter and Larisa Ponomareva. 2019. Komi latin letters, degrees of UNICODE facilitation. *Proceedings of the Language Technologies for All (LT4All)*.

Jack Michael Rueter. 1995. *Komia-angliskö̈j-finskoj = Komi-English-Finnish = Komilais-englantilais-suomalainen*. Self-published.

V. Tauli. 1956. The origin of affixes. *Finnisch-ugrische Forschungen*, XXXII(Heft 1–2):170–225.

Jörg Tiedemann. 2021. The development of a comprehensive data set for systematic studies of machine translation. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*, pages 248–262. Rootroo Ltd., Helsinki. This book has been authored for Jack Rueter in honor of his 60th birthday.

Trond Trosterud. 2004. Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92.

Trond Trosterud. 2004b. Porting morphological analysis and disambiguation to new languages. In *Poster presented at SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*.

Jussi Ylikoski. 2020. Kielemme kääpiösijoista: prolatiivi, temporaali ja distributiivi. *Virittäjä*, (4):529–554.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė,

Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Р Блокланд, М Рисслер, Н Партанен, А Чемышев, and М Федина. 2014. Использование цифровых корпусов и компьютерных программ в диалектологических исследованиях: теория и практика. In *Актуальные проблемы диалектологии языков народов России: материалы XIV всеросс. науч. конф., посвященной*, pages 20–22.

В.И. Лыткин and Д.А. Тимушев. 1961. *Коми-*

русский словарь. Государственное издательство уностранных и национальных словарей, Москва.

Н.Д. Манова. 1994. *Учимся говорить по-коми. Самоучитель коми языка*. Коми книжное издательство, Сыктывкар.

Г. Некрасова. 2000. Эмакыв. In Г. В. Федюнёва, editor, *Öнія коми кыв, морфология*. Россияса наукаяс академия, Коми наука шӧрин, Сыктывкар.

Марина Серафимовна Федина. 2019. Корпус коми языка как база для научных исследований. In *II Международная научная конференция «Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы» проводится в рамках реализации Государственной программы «Сохранение и развитие государственных языков Республики Башкортостан и языков народов Республики Башкортостан» на 2019–2024 гг. Ответственный редактор: Ахмадеева АУ*, page 45.

Евгений Александрович Цыпанов. 1992. *Коми кыв: самоучитель коми языкаю*. Коми кн. изд-во, Сыктывкар.

Йӧлгинь Цыпанов. 2021. Питирим Сорокинлысь «a long journey» небӧг комиӧдӧмын шыбӧльяс. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*, pages 94–103. Rootroo Ltd., Helsinki. This book has been authored for Jack Rueter in honor of his 60th birthday.

Василий Пантелеймонович Чупров. 2018. Электронный корпус ижемского диалекта коми языка как ресурс для исследования речи ижемских коми. In *Говоры Республики Коми и сопредельных областей*, pages 158–170.

Öньӧ Лав. 2015. Видзам-сӧвмӧдам коми кыв! *Арт*, (3):135–144.